Summarizing Egocentric Video

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Yong Jae Lee and Lu Zheng

THE UNIVERSITY OF TEXAS AT AUSTIN

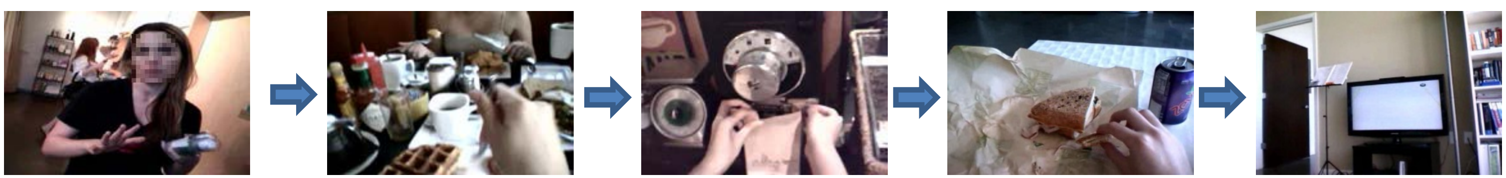

~1990
M.I.T.
Steve Mann


2013

Goal: Summarize egocentric video

Wearable camera

Input: Egocentric video of the camera wearer's day



9:00 am 10:00 am 11:00 am 12:00 pm 1:00 pm 2:00 pm

Output: Storyboard (or video skim) summary

Potential applications of egocentric video summarization

Memory aid

Law enforcement

Mobile robot discovery

RHex Hexapedal Robot, Penn's GRASP Laboratory

What makes egocentric data hard to summarize?

- Subtle event boundaries
- Subtle figure/ground
- Long streams of data

Prior work

- **Egocentric recognition**

  [Starner et al. 1998, Doherty et al. 2008, Spriggs et al. 2009, Jojic et al. 2010, Ren & Gu 2010, Fathi et al. 2011, Aghazadeh et al. 2011, Kitani et al. 2011, Pirsiavash & Ramanan 2012, Fathi et al. 2012,…]

- **Video summarization**

  [Wolf 1996, Zhang et al. 1997, Ngo et al. 2003, Goldman et al. 2006, Caspi et al. 2006, Pritch et al. 2007, Laganiere et al. 2008, Liu et al. 2010, Nam & Tewfik 2002, Ellouze et al. 2010,…]

→ Low-level cues, stationary cameras

→ Consider summarization as a *sampling* problem

Our idea:
Story-driven summarization

[Lu & Grauman, CVPR 2013]

Our idea: Story-driven summarization

Good summary captures the progress of the story

1. Segment video temporally into subshots
2. Select chain of *k* subshots that maximize both weakest link's influence and object importance

[Lee & Grauman, CVPR 2012; Lu & Grauman, CVPR 2013]

Egocentric subshot detection

Define 3 generic ego-activities:

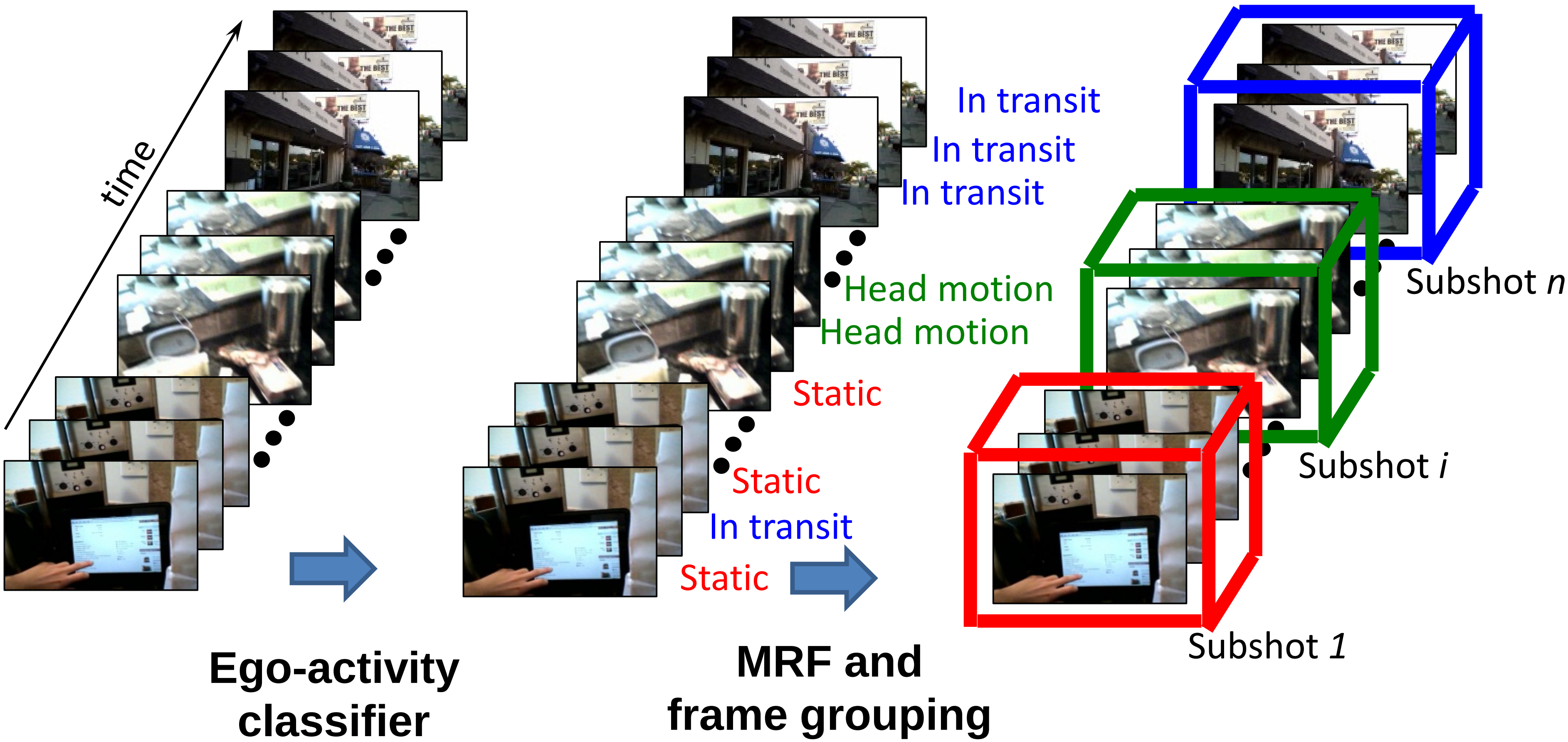
~Static

In transit

Head moving

- Train classifiers to predict these activity types
- Features based on flow and motion blur

Egocentric subshot detection



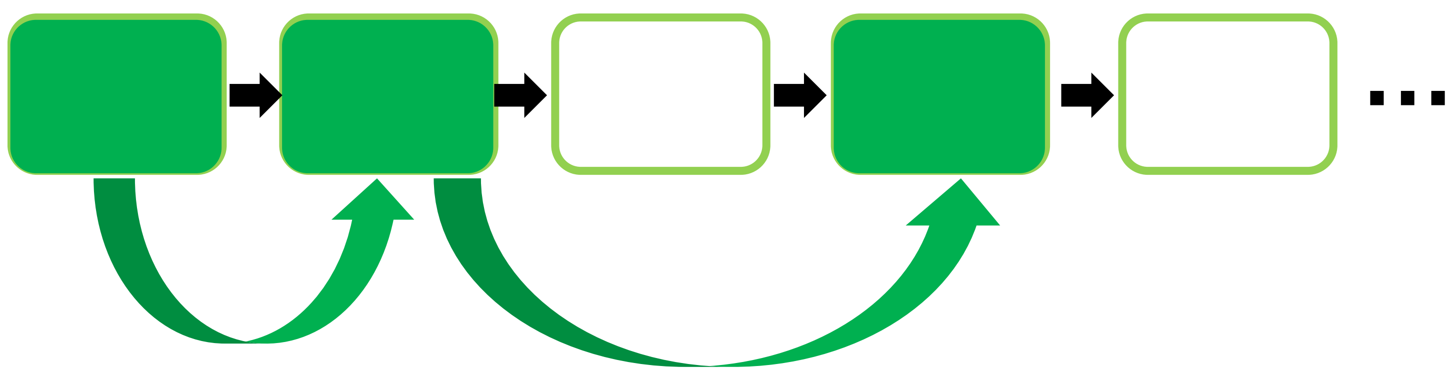
Subshot selection objective

Good summary = chain of *k* selected subshots in which each influences the next via some subset of key objects

$$S^* = \arg\max_{S \subset \mathcal{V}} \; \lambda_s \, \mathcal{S}(S) + \lambda_i \, \mathcal{I}(S) + \lambda_d \, \mathcal{D}(S)$$

influence **importance** **diversity**

Subshots



Learning region importance

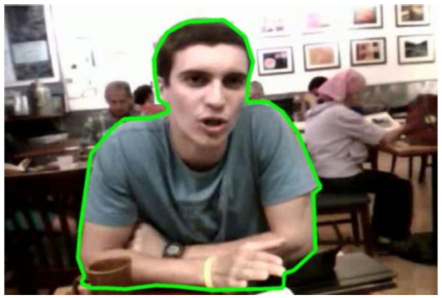
Man wearing a blue shirt and watch in coffee shop

Yellow notepad on table

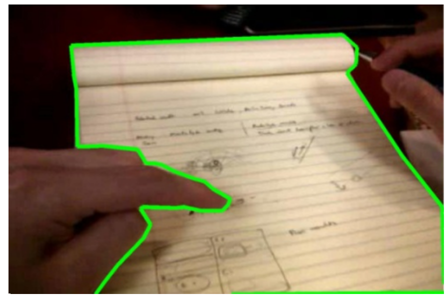
Coffee mug that cameraman drinks

- First task: watch a short clip, and *describe in text* the essential people or objects necessary to create a summary

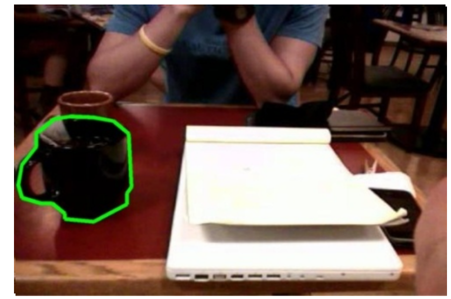
Learning region importance



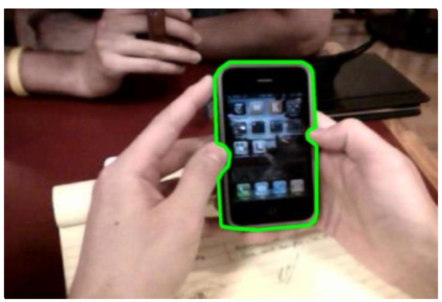
Man wearing a blue shirt and watch in coffee shop



Yellow notepad on table



Coffee mug that cameraman drinks



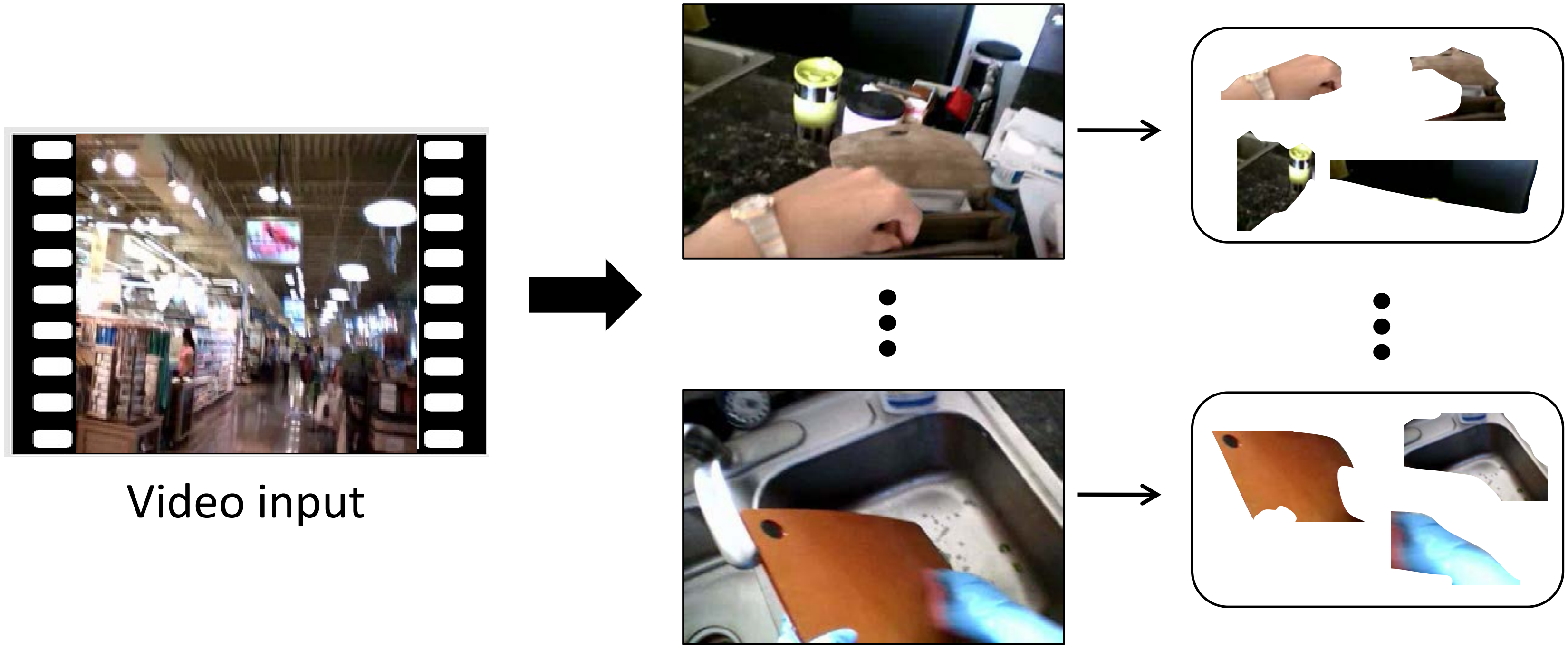
Iphone that the camera wearer holds

Camera wearer cleaning the plates

Soup bowl

- Second task: draw polygons around any described person or object *obtained from the first task* in sampled frames

Learning region importance



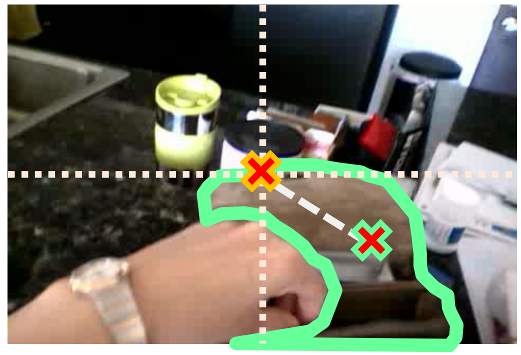
Video input

Generate candidate object regions
for uniformly sampled frames

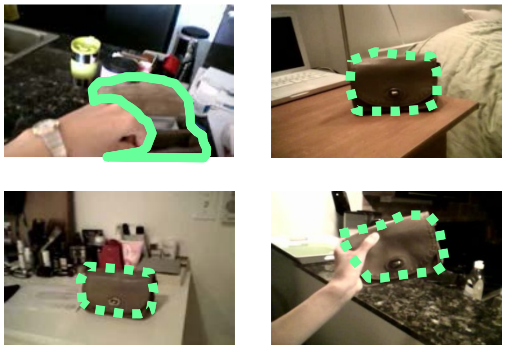
Learning region importance

Egocentric features:

distance to hand



distance to frame center



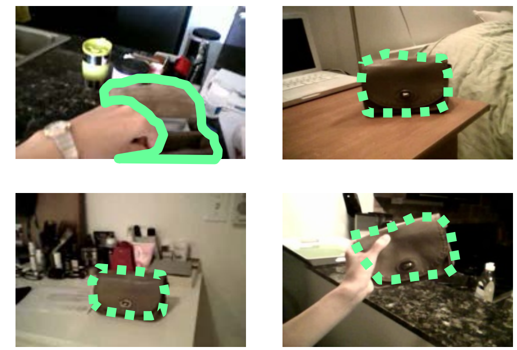
frequency

Learning region importance

Egocentric features:

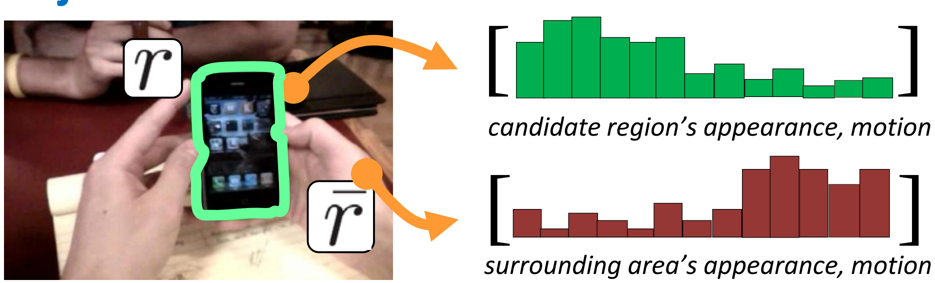
distance to hand

distance to frame center



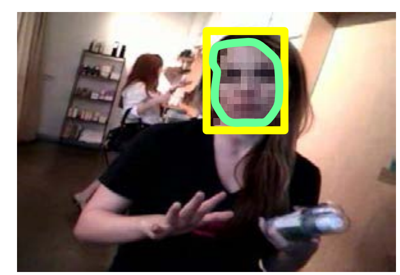
frequency

Object features:



"Object-like" appearance, motion

[Endres et al. ECCV 2010, Lee et al. ICCV 2011]



overlap w/ face detection

Region features: *size, width, height, centroid*

Learning region importance

$$I(r) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i(r) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \beta_{i,j} x_i(r) x_j(r)$$

importance learned parameters i'th feature value

- Regressor to predict a region's *degree* of importance
- Expect significant interactions between the features
- For training: $I(r) = \frac{|GT \cap r|}{|GT \cup r|}$
- For testing: predict $I(r)$ given $x_i(r)$'s

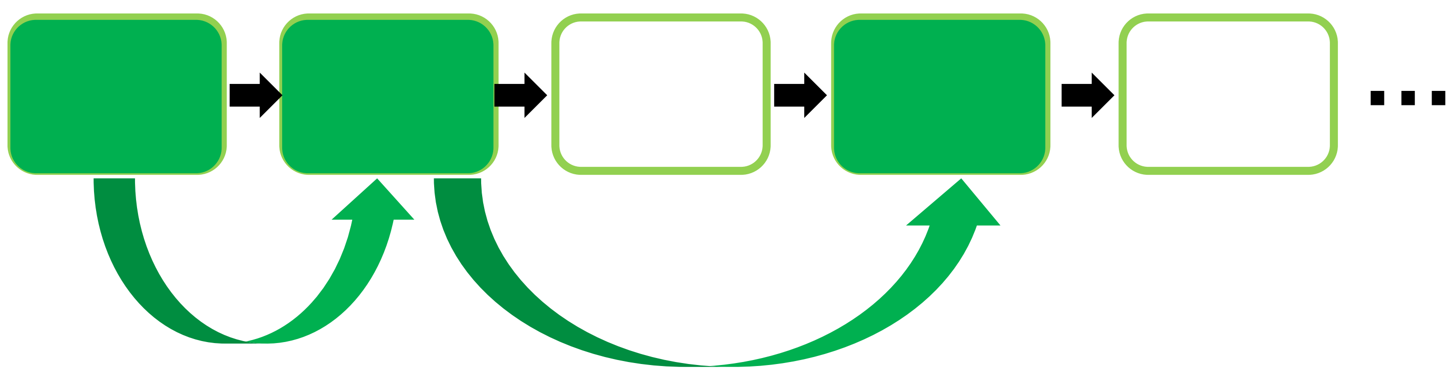
Subshot selection objective

Good summary = chain of *k* selected subshots in which each influences the next via some subset of key objects

$$S^* = \arg\max_{S \subset \mathcal{V}} \; \lambda_s \, \mathcal{S}(S) + \lambda_i \, \mathcal{I}(S) + \lambda_d \, \mathcal{D}(S)$$

influence **importance** **diversity**

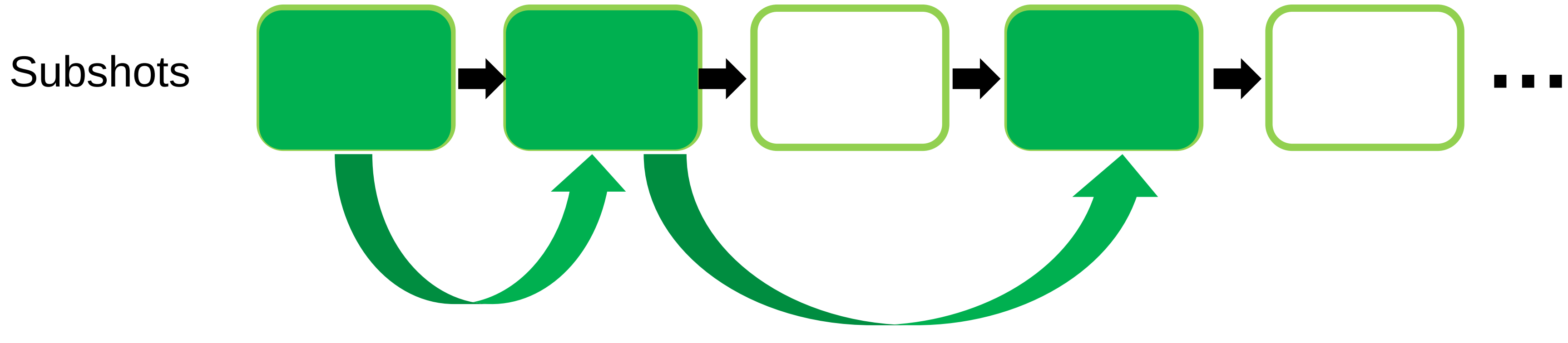
Subshots



Influence criterion

- Want the *k* subshots that maximize the weakest link's influence, subject to coherency constraints

$$\mathcal{S}(S) = \max_{a} \min_{j=1,\ldots,K-1} \sum_{o_i \in O} a_{i,j} \textsc{Influence}(s_j, s_{j+1} | o_i)$$



Document-document influence
[Shahaf & Guestrin, KDD 2010]

CNNMoney
A Service of CNN, Fortune & Money
FORTUNE Money
Home | Video | Business News | Markets | Term Sheet | Economy | Tech | Personal Finance

REAL ESTATE
Mortgage Meltdown
Archive

Home prices post record decline

S&P/Case-Shiller index of 10 major cities fell 6.7% in October. Housing markets remain 'grim.'

By Les Christie, CNNMoney.com staff writer
December 26 2007: 3:03 PM EST

EMAIL | PRINT | DIGG | RSS

NEW YORK (CNNMoney.com) -- Home prices fell 6.7 percent in October, compared with a year ago, according to the S&P/Case-Shiller 10-city home-price index. It was the largest drop recorded since the index began in 1987.

It marked the 10th consecutive month of price depreciation and 23 months of decelerating returns.

Special Report
FORECLOSURE
MORTGAGE MELTDOWN
- Seniors face grim choices amid

CNN Politics
Justice | Entertainment | Tech | Health | Living | Travel | Opinion | iReport | Money

HEALTH CARE

Health-care debate heats up as Senate, House grapple with plans

June 08, 2009

Share | Twitter | Email
Recommend | 44 people recommend this

As the debate on health-care reform heats up on Capitol Hill, it's clear lawmakers don't see eye-to-eye on the issue -- with each other or President Obama.

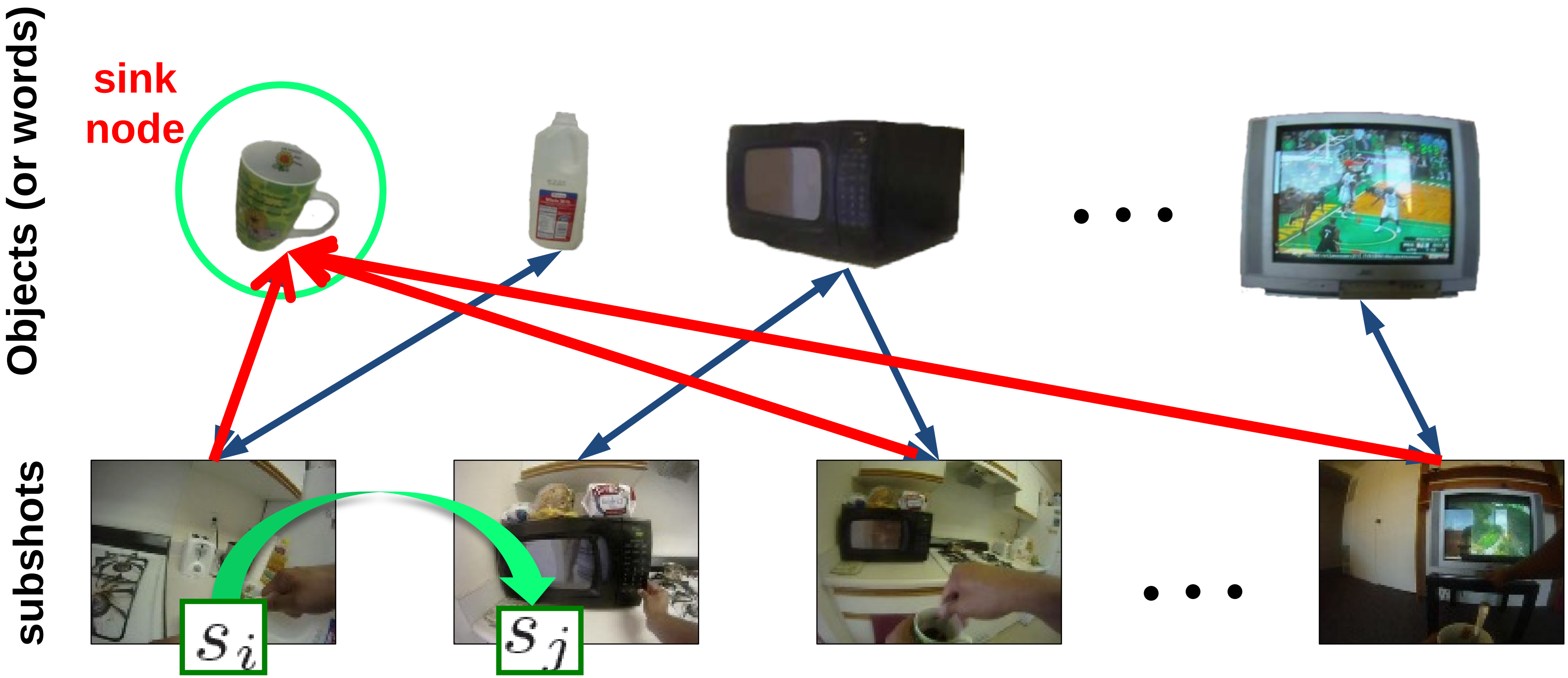
Obama told Congress this past weekend that it's time to deliver on health-care reform, and he wants a bill on his desk by October at the latest. But this week already is demonstrating just how difficult and complex coming up with a nuts-and-bolts bill is.

In the Senate, key negotiators broke up a session Monday still stuck on whether to create a government-run health-insurance plan to compete with private insurers -- something Obama and most Democrats want, and most Republicans oppose.

President Obama says a public health pla
consumers and keep costs down.

Connecting the dots between news articles. D. Shahaf and C. Guestrin. In KDD, 2010.

Estimating visual influence

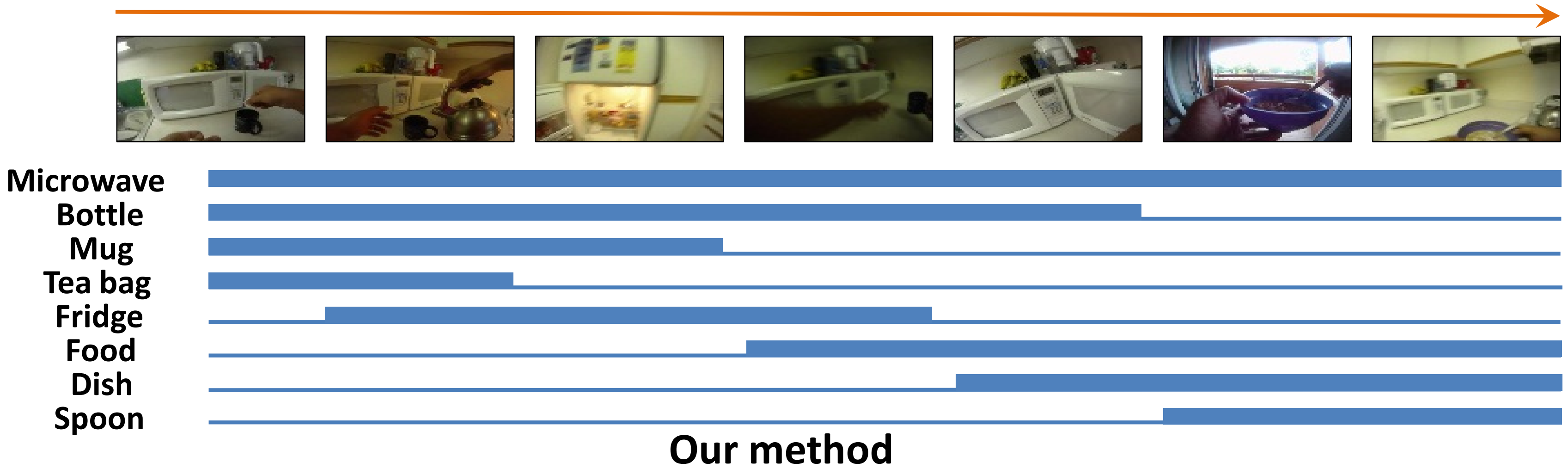


$$\textsc{Influence}(s_i, s_j | o) = \prod_i (s_j) - \prod_i^o (s_j)$$

Captures how reachable subshot *j* is from subshot *i*, via any object *o*

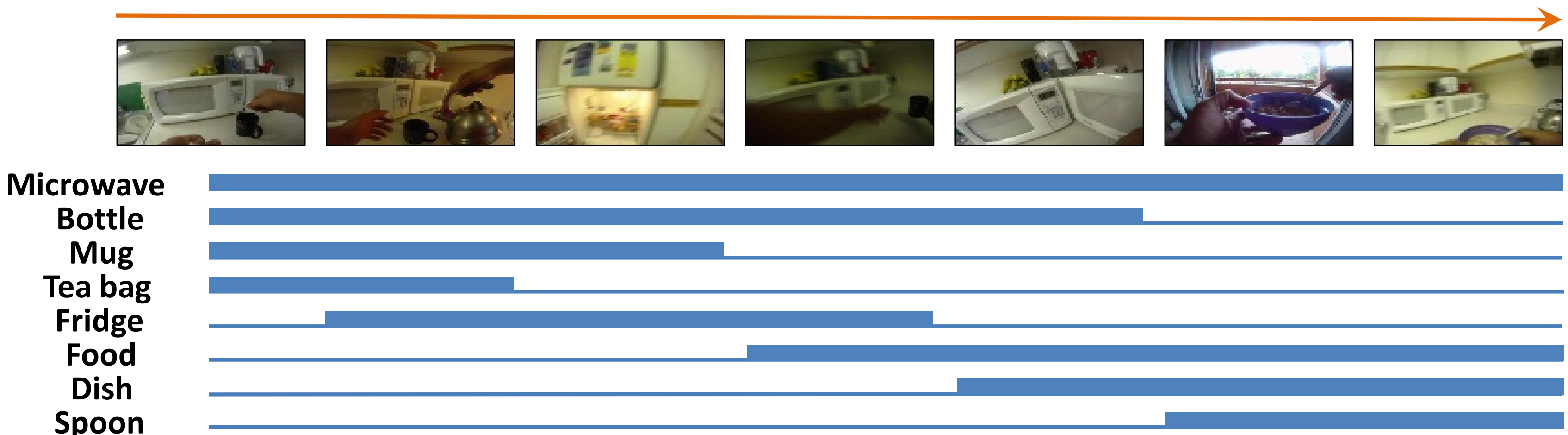
Estimating visual influence

- Prefer small number of objects at once, and coherent (smooth) entrance/exit patterns

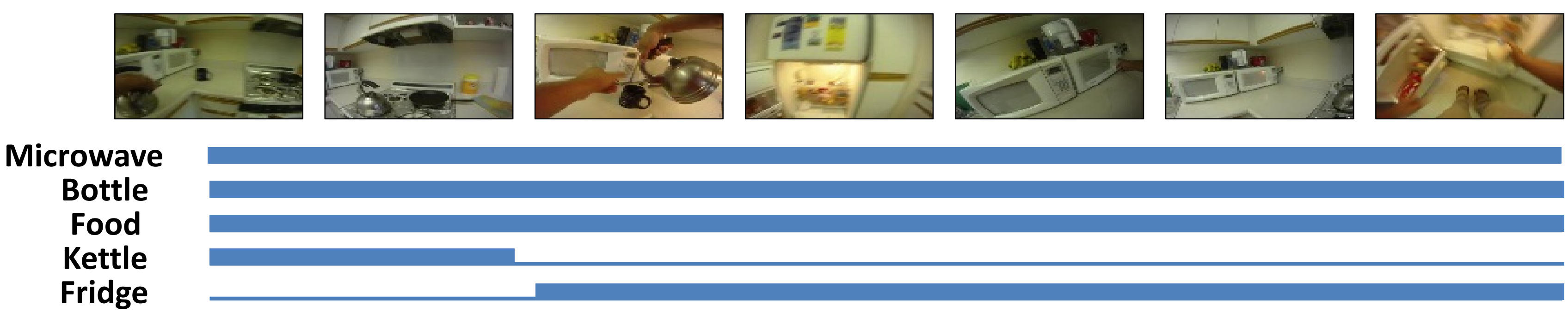


Estimating visual influence

- Prefer small number of objects at once, and coherent (smooth) entrance/exit patterns



Our method



Uniform sampling

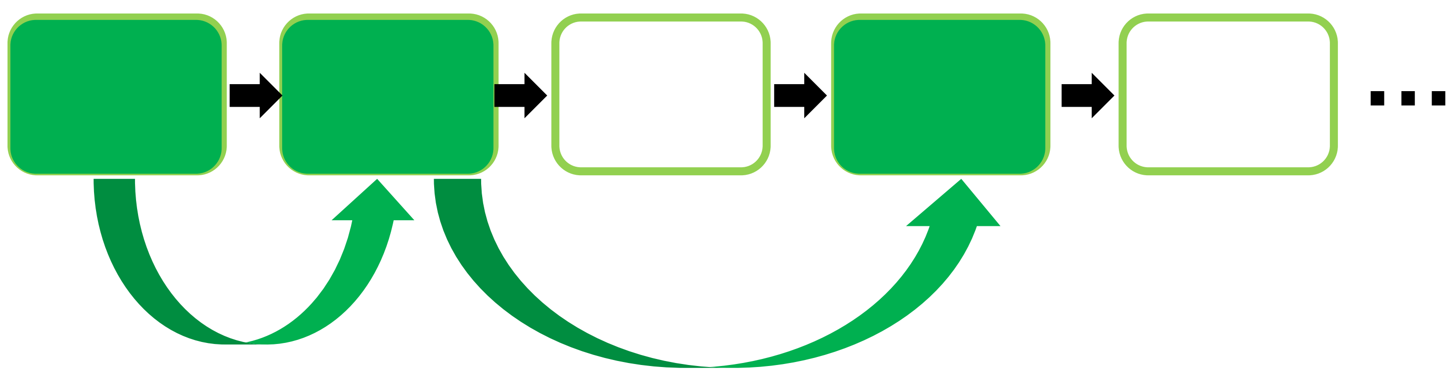
Subshot selection objective

Good summary = chain of *k* selected subshots in which each influences the next via some subset of key objects

$$S^* = \arg\max_{S \subset \mathcal{V}} \; \lambda_s \, \mathcal{S}(S) + \lambda_i \, \mathcal{I}(S) + \lambda_d \, \mathcal{D}(S)$$

influence **importance** **diversity**

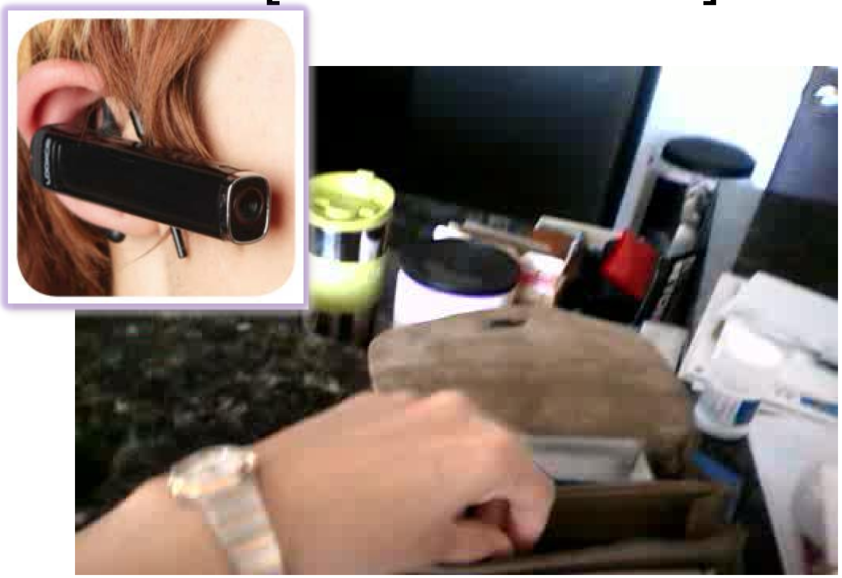
Subshots



Optimize with aid of priority queue of (sub)-chains

Datasets

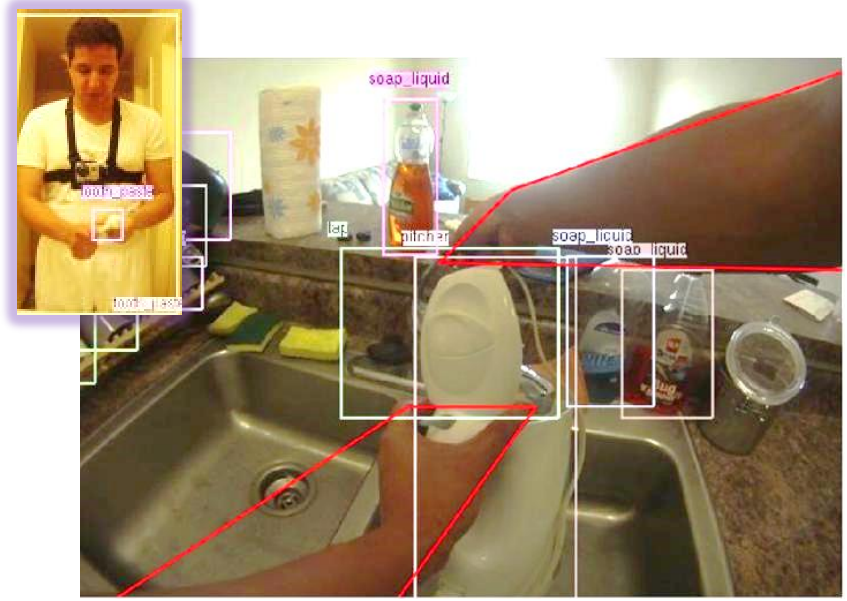
UT Egocentric (UTE)
[Lee et al. 2012]



4 videos, each 3-5 hours long, uncontrolled setting.

We use visual words and subshots.

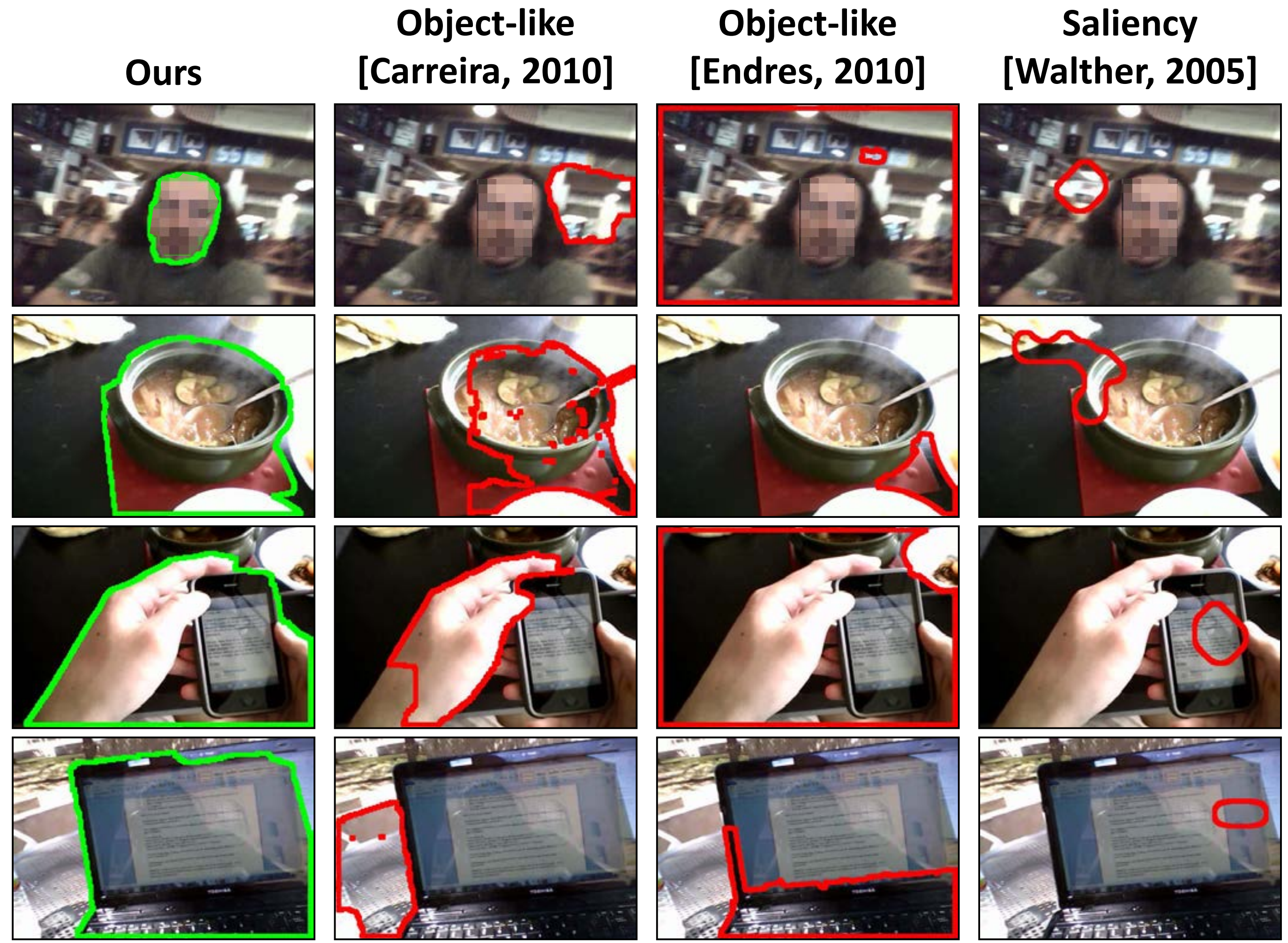
Activities of Daily Living (ADL)
[Pirsiavash & Ramanan 2009]



20 videos, each 20-60 minutes, daily activities in house.

We use object bounding boxes and keyframes.

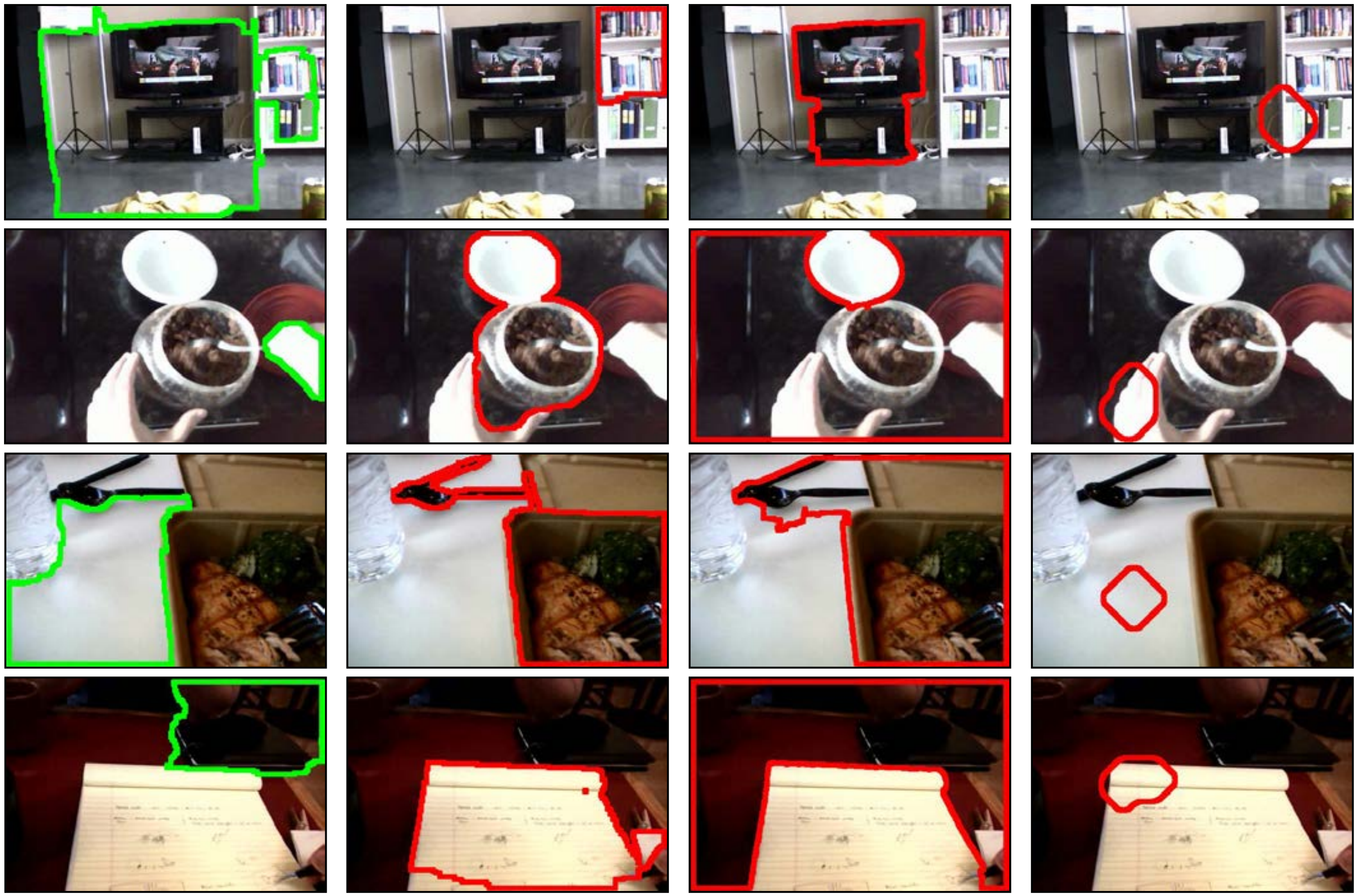
Results: Important region prediction



Good predictions

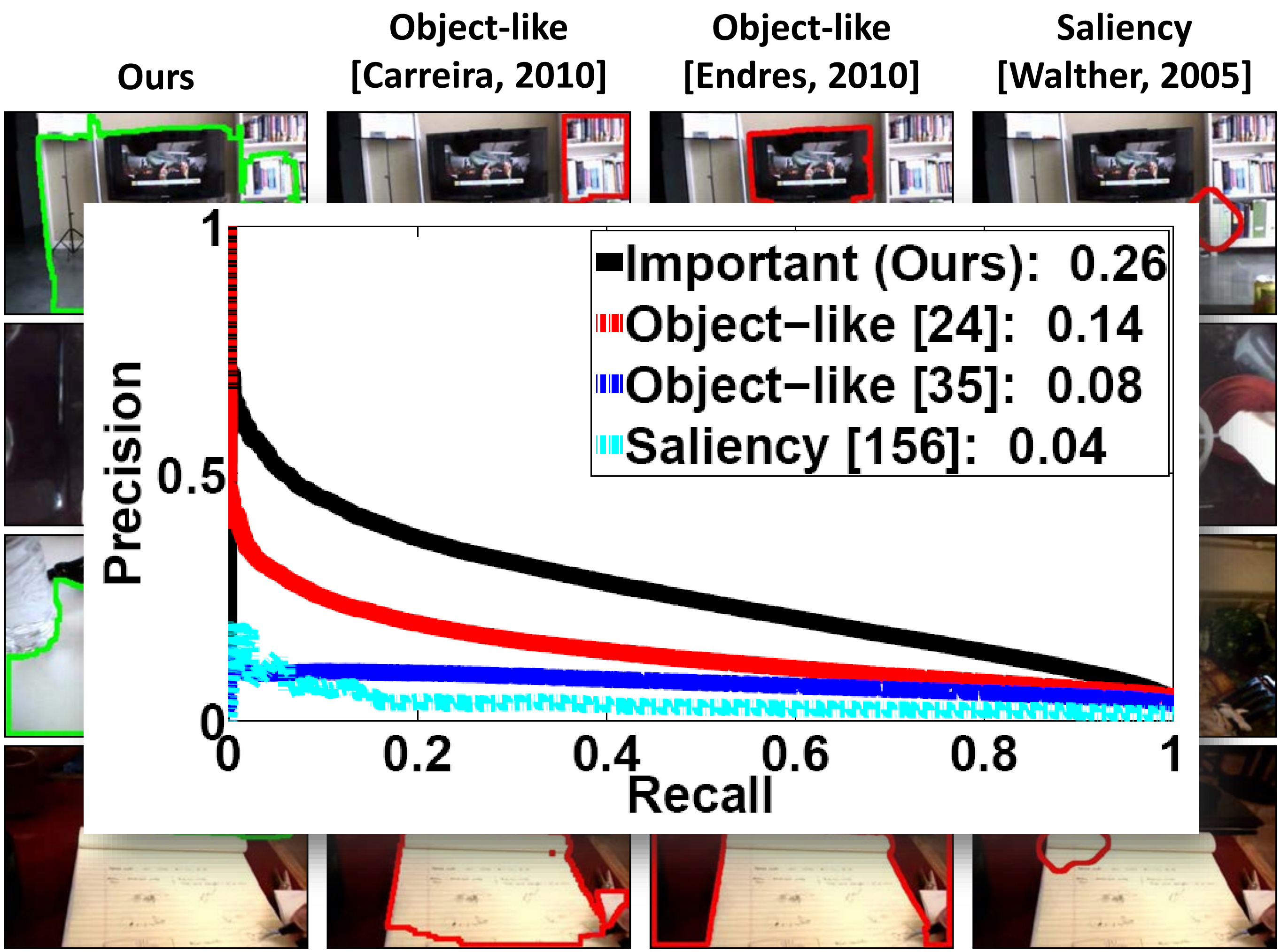
Results: Important region prediction

Ours	Object-like [Carreira, 2010]	Object-like [Endres, 2010]	Saliency [Walther, 2005]



Failure cases

Results: Important region prediction



Failure cases

Example keyframe summary – UTE data

Original video (3 hours)

Our summary (12 frames)

Example keyframe summary – UTE data

Alternative methods for comparison

Uniform keyframe sampling
(12 frames)

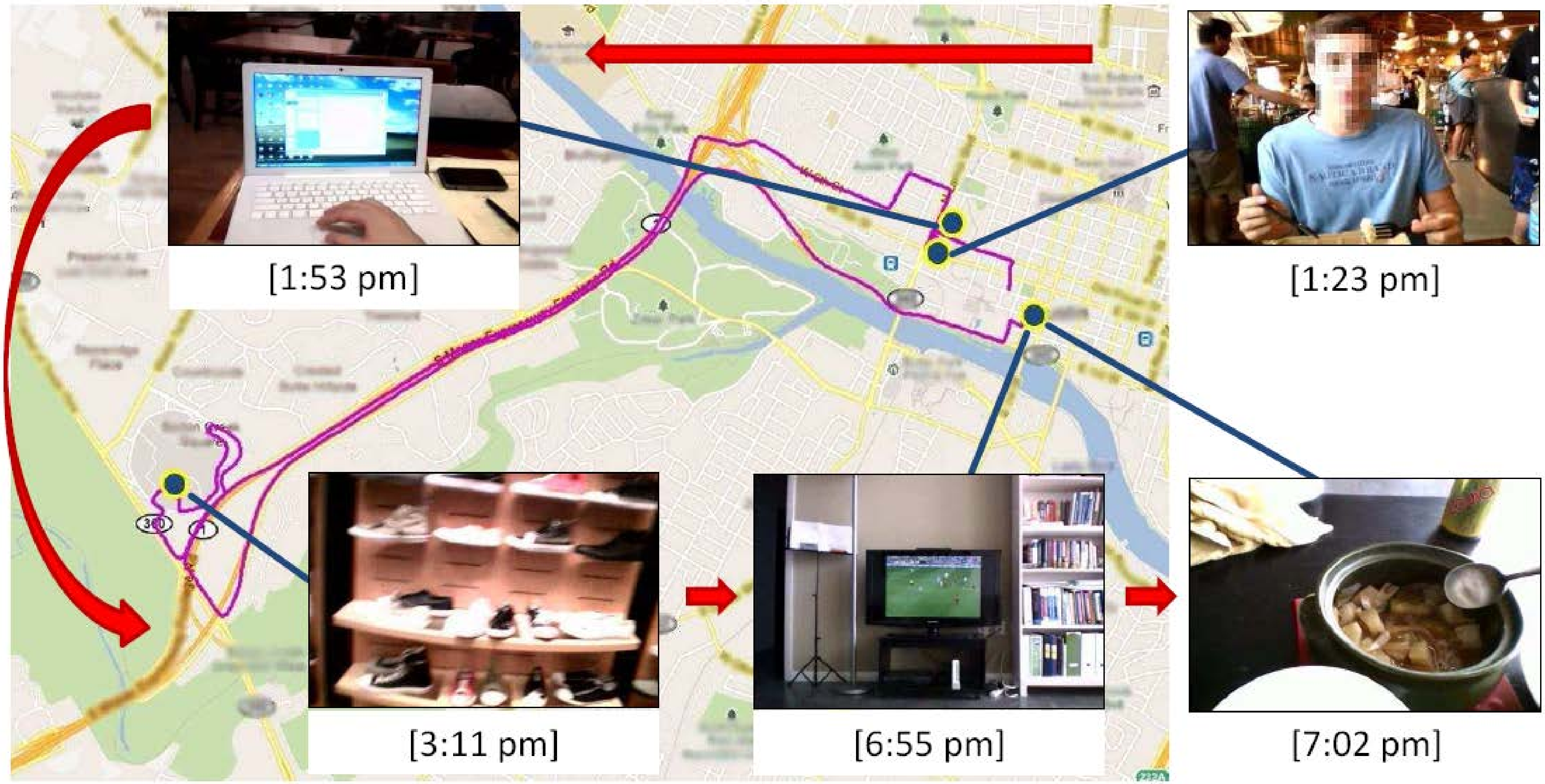
[Liu & Kender, 2002]
(12 frames)

Example summary – UTE data

Ours

Baseline

Generating storyboard maps



Augment keyframe summary with geolocations

[Lee & Grauman, CVPR 2012]

How to evaluate a summary?

- Blind taste tests: which better captures…?
  - Your real-life experience (camera wearer)
  - This text description you read
  - The sped up original video you watched

- Compared methods:
  - Uniform sampling
  - Shortest path on subshots' object similarity
  - Importance-driven summaries (Lee et al. 2012)
  - Event-detection followed by sampling
  - Diversity-based objective (Liu & Kender 2002)

Human subject results: Blind taste test

How often do subjects prefer our summary?

Data	Uniform sampling	Shortest-path	Object-driven Lee et al. 2012
UTE	**90.0%**	**90.9%**	**81.8%**
ADL	**75.7%**	**94.6%**	**N/A**

34 human subjects, ages 18-60
12 hours of original video
Each comparison done by 5 subjects

Total 535 tasks, 45 hours of subject time

Next steps

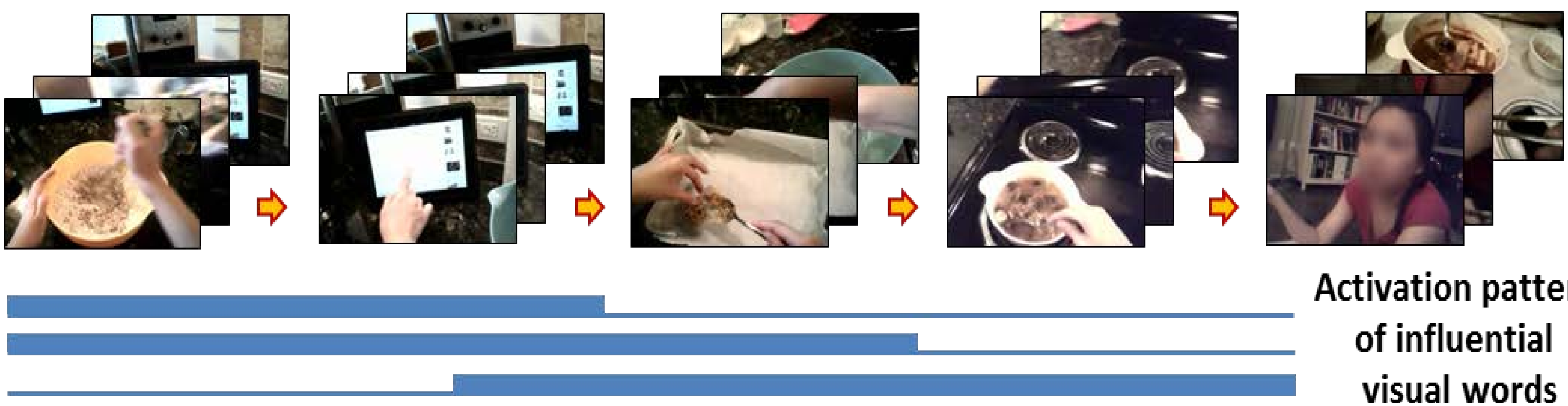
- Summaries while streaming
- Multiple scales of influence
- Object-centric → activity-centric?
- Additional sensors
- Evaluation as an explicit index

Summary

- Have more video than can be watched!

  → Need **summaries** to access and browse

- First person story-driven video summarization
  - Egocentric temporal segmentation
  - Estimate influence between events given their objects
  - Category-independent region importance prediction



Activation pattern of influential visual words

References

- Discovering Important People and Objects for Egocentric Video Summarization. Y. J. Lee, J. Ghosh, and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 2012.

- Story-Driven Summarization for Egocentric Video. Z. Lu and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 2013.